Align Deep Features for Oriented Object Detection

Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia, *Senior Member, IEEE*

***Abstract*— The past decade has witnessed significant progress on detecting objects in aerial images that are often distributed with large-scale variations and arbitrary orientations. However, most of existing methods rely on heuristically defined anchors with different scales, angles, and aspect ratios, and usually suffer from severe misalignment between anchor boxes (ABs) and axis-aligned convolutional features, which lead to the common inconsistency between the classification score and localization accuracy. To address this issue, we propose a *single-shot alignment network* (S^2A-Net) consisting of two modules: a feature alignment module (FAM) and an oriented detection module (ODM). The FAM can generate high-quality anchors with an anchor refinement network and adaptively align the convolutional features according to the ABs with a novel alignment convolution. The ODM first adopts active rotating filters to encode the orientation information and then produces orientation-sensitive and orientation-invariant features to alleviate the inconsistency between classification score and localization accuracy. Besides, we further explore the approach to detect objects in large-size images, which leads to a better trade-off between speed and accuracy. Extensive experiments demonstrate that our method can achieve the state-of-the-art performance on two commonly used aerial objects' data sets (i.e., DOTA and HRSC2016) while keeping high efficiency.**

***Index Terms*— Aerial images, deep learning, feature alignment, object detection.**

I. Introduction

OBJECT detection in aerial images aims at identifying the locations and categories of objects of interest (e.g., planes, ships, vehicles). With the framework of deep convolutional neural networks, object detection in aerial images (ODAI) has made significant progress in recent years [1]–[7], where most of existing methods are devoted to cope with the challenges raised by the large-scale variations and arbitrary orientations of crowded objects in aerial images.

Manuscript received August 21, 2020; revised January 9, 2021; accepted February 6, 2021. Date of publication March 12, 2021; date of current version December 9, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61922065, Grant 61771350, Grant 41820104006, and Grant 61871299; in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170; and in part by the Shanghai Aerospace Science and Technology Innovation Project under Grant SAST2019-094. *(J. Han and J. Ding contributed equally to this work.) (Corresponding author: Gui-Song Xia.)*

Jiaming Han and Jian Ding are with the State Key Laboratory LIESMARS, Institute of Artificial Intelligence, Wuhan University, Wuhan 430079, China.

Jie Li is with the Shanghai Aerospace Electronic Technology Institute, Shanghai 201108, China (e-mail: trackerdsp@163.com).

Gui-Song Xia is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Institute of Artificial Intelligence, Wuhan University, Wuhan 430072, China, and also with the State Key Laboratory LIESMARS, Wuhan University, Wuhan 430072 China (e-mail: guisong.xia@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3062048

To achieve better detection performance, most state-of-the-art aerial object detectors [4], [5], [7], [8] rely on the complex R-CNN [9] frameworks, which consist of two parts: a region proposal network (RPN) and an R-CNN detection head. In a general pipeline, RPN is used to generate high-quality region of interests (RoIs) from horizontal anchors, then an RoI pooling operator is adopted to extract accurate features from RoIs. Finally, R-CNN is employed to regress the bounding boxes and classify them. However, it is worth noticing that horizontal RoIs often result in severe misalignment between bounding boxes and oriented objects [3], [4]. For example, a horizontal RoI usually contains several instances due to oriented and densely packed objects in aerial images. A natural solution is employing oriented bounding boxes as anchors to alleviate this issue [2], [3]. As a consequence, well-designed anchors with different angles, scales, and aspect ratios are required, which however leads to massive computations and memory footprint. Recently, RoI transformer [4] was proposed to address this issue by transforming horizontal RoIs into rotated RoIs, avoiding a large number of anchors, but it still needs heuristically defined anchors and complex RoI operation.

In contrast with R-CNN-based detectors, one-stage detectors regress the bounding boxes and classify them directly with regular and densely sampling anchors. This architecture enjoys high-computational efficiency but often lags behind in accuracy [3]. As shown in Fig. 1(a), we argue that severe misalignment in one-stage detectors matters.

1) Heuristically defined anchors are with low-quality and cannot cover the objects, leading a misalignment between objects and anchors. For example, the aspect ratio of a bridge usually ranges from 1/3 to 1/30, and only a few or even no anchors can be assigned to it. This misalignment usually aggravates the foreground–background class imbalance and hinders the performance.
2) The convolutional features from the backbone network are usually axis-aligned with fixed receptive field, while objects in aerial images are distributed with arbitrary orientations and variant appearances. Even an AB is assigned to an instance with high confidence [i.e., Intersection over Union (IoU)], there is still a misalignment between ABs and convolutional features. In other words, the corresponding feature of an anchor box is hard to represent the whole object to some extent. As a result, the final classification score cannot accurately reflect the localization accuracy, which also hinders the detection performance in post-processing phases [e.g., non-maximum suppression (NMS)].

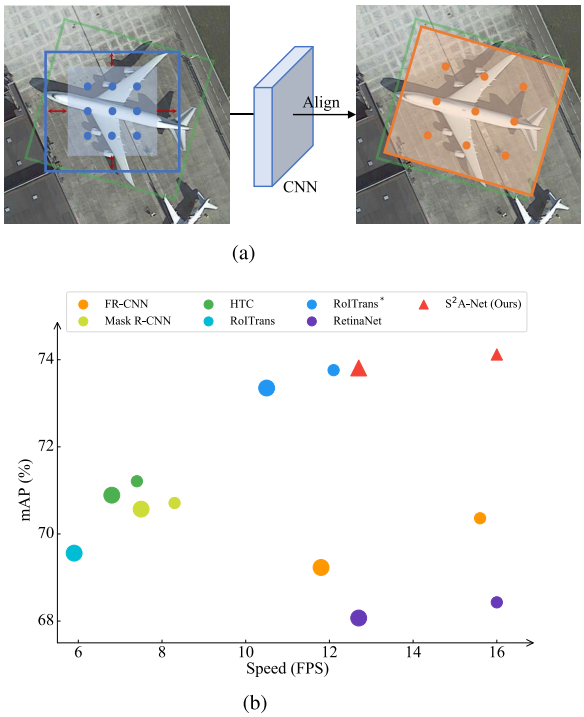


Fig. 1. (a) Misalignment (red arrows) between an AB (blue bounding box) and convolutional features (light blue rectangle). To alleviate this issue, we first refine the initial anchor into a rotated one (orange bounding box), and then adjust the feature sampling locations (orange points) with the guide of the refined AB to extract aligned deep features. The green box denotes the ground truth. (b) Performance comparisons of different methods under the same settings: ResNet50 (in small markers) and ResNet101 (in big markers) backbones, 1024×1024 input size of images, without data augmentation. Faster R-CNN (FR-CNN) [10], Mask R-CNN [11], RetinaNet [12], Hybrid Task Cascade (HTC) [13], and RoI Transformer (RoITrans) [4] are tested. The speed of all methods is reported on the V100 GPU in terms of FPS. It is noted that Mask R-CNN, HTC, and RoITrans are tested based on the `AerialDetection`[1] project. RoITrans* indicates an official re-implementation.

To address these issues in one-stage detectors, we propose a *Single-Shot Alignment Network* (S^2A-Net) which consists of two modules: a feature alignment module (FAM) and an oriented detection module (ODM). The FAM can generate high-quality anchors with an anchor refinement network (ARN) and adaptively align the feature according to the corresponding ABs [Fig. 1(a)] with an alignment convolution (AlignConv). Different from other methods with densely sampling anchors, we employ only one squared anchor for each location in the feature map, and the ARN refines them into high-quality rotated anchors. Then the AlignConv, a variant of convolution, adaptively aligns the feature according to the shapes, sizes, and orientations of its corresponding anchors.[1] In the ODM, we first adopt active rotating filters (ARF) [14] to encode the orientation information and produce orientation-sensitive features, and then extract orientation-invariant features by pooling the orientation-sensitive features. Finally, we feed the features into a regression sub-network and a classification sub-network to yield the final predictions. Besides, we also explore the approach to detect objects on large-size images (e.g., 4000×4000) rather than on chip images, which significantly reduces the overall inference time with negligible loss of accuracy. Extensive experiments on commonly used data sets, i.e., DOTA [3] and HRSC2016 [15], demonstrate that our proposed method can achieve state-of-the-art performance while keeping high efficiency, see Fig. 1(b).

Our main contributions are summarized as follows.

1) We propose a novel alignment convolution to alleviate the misalignment between axis-aligned convolutional features and arbitrary oriented objects in a fully convolutional way. It is noted that AlignConv has negligible extra consuming time compared with standard convolution and can be embedded into many detectors with little modification.
2) With the alignment convolution embedded, we design a light single-shot alignment network which enables us to generate high-quality anchors and aligned features for accurate object detection in aerial images.
3) We report 79.42% mAP on the oriented object detection task on the DOTA data set, achieving the state-of-the-art in both speed and accuracy.

The rest of this article is organized as follows. Section II introduces the related works. Section III introduces the details of our proposed S^2A-Net. In Section IV, the experimental results and analysis are reported on challenging DOTA and HRSC2016 data sets. Finally, the conclusion is made in Section V.

II. Related Works

With the advance of machine learning, especially deep learning, object detection has made significant progress in recent years, which can be roughly divided into two groups: two-stage detectors and one-stage detectors. Two-stage detectors [9]–[11], [16] first generate a sparse set of RoIs in the first stage, and perform an RoI-wise bounding box regression and object classification in the second one. One-stage detectors, e.g., YOLO [17] and SSD [18], detect objects directly and do not require the RoI generation stage. Generally, the performance of one-stage detectors usually lag behind two-stage detectors due to extreme foreground–background class imbalance. To address this problem, the *Focal Loss* [12] can be used, and anchor-free detectors [19]–[21] alternatively formulate object detection as a points detection problem to avoid complex computations related to anchors and usually run faster.

A. Object Detection in Aerial Images

Objects in aerial images are often crowded, distribute with large-scale variations and appear at arbitrary orientations. Generic object detection methods with horizontal anchors [3] usually suffer from severe misalignment in such scenarios: one anchor/RoI may contain several instances. Some methods [2], [22], [23] adopt rotated anchors with different angles, scales, and aspect ratios to alleviate this issue, while involving heavy computations related to anchors (e.g., bounding box transform and groundtruth matching). Ding *et al.* [4] proposed RoI Transformer to transform horizontal RoIs into rotated RoIs,

[1]https://github.com/dingjiansw101/AerialDetection

which avoids a large number of anchors and alleviates the misalignment issue. However, it still needs heuristically defined anchors and complex RoI operations. Instead of employing rotated anchors, Xu *et al.* [7] glide the vertex of the horizontal bounding box to accurately describe an oriented object. But the corresponding feature of a RoI is still horizontal and suffers from the misalignment issue. Recently proposed R^3Det [24] samples features from five locations (e.g., center and corners) of the corresponding AB and sum them up to re-encode the position information. In contrast with the above methods, the proposed S^2A-Net in this article gets ride of heuristically defined anchors and can generate high-quality anchors by refining horizontal anchors into rotated anchors. Besides, the proposed FAM module enables to achieve feature alignment in a fully convolutional way.

B. Feature Alignment in Object Detection

Feature alignment usually refers to the alignment between convolution features and ABs/RoIs, which is important for both two-stage and one-stage detectors. Detectors relying on misaligned features are hard to obtain accurate detections. In two-stage detectors, an RoI operator (e.g., RoIPooling [16], RoIAlign [11], and deformable RoIPooling [25]) is adopted to extract fixed-length features inside the RoIs which can approximately represent the location of objects. RoIPooling first divides a RoI into a grid of sub-regions and then max-pools each sub-region into the corresponding output grid cell. However, RoIPooling quantizes the floating-number boundary of a RoI into integer, which introduces misalignment between the RoI and the feature. To avoid the quantization of RoIPooling, RoIAlign adopts bilinear interpolation to compute the extract values at each sampling location in sub-regions, significantly boosting the performance of localization. Meanwhile, deformable RoIPooling adds an offset to each sub-region of a RoI, enabling adaptive feature selection. However, the RoI operator usually involves massive region-wise operation, e.g., feature warping and feature interpolation, which becomes a bottleneck toward fast object detection.

Recently, guided anchoring [26] tries to align features with the guide of anchor shapes. It learns an offset field from the anchor prediction map and then guides the deformable convolution (DeformConv) to extract aligned features. AlignDet [27] designs an RoI Convolution to obtain the same effect as RoIAlign in one-stage detector. Both [26] and [27] achieve feature alignment in a fully convolutional way and enjoy high efficiency. These methods work well for objects in nature images but often lose their performance when detecting objects that are oriented and densely packed in aerial images, although some of them (e.g., Rotated RoIPooling [23] and rotated position sensitive RoIAlign [4]) have been adopted to achieve feature alignment in oriented object detection. Different from the aforementioned methods, our proposed method aims at alleviating the misalignment between axis-aligned convolutional features and arbitrary-oriented objects, which adjusts the feature sampling locations with the guide of ABs.

C. Inconsistency Between Regression and Classification

An object detector usually consists of two parallel tasks: bounding-box regression and object classification, which share the same features from the backbone network. And the classification score is used to reflect the localization accuracy in a post-processing phase (e.g., NMS). However, as discussed in [28] and [29], there is a common inconsistency between classification score and localization accuracy. Detections with high classification scores may produce bounding boxes with low localization accuracy, while other nearby detections with high localization accuracy may be suppressed in the NMS step. To address this issue, IoU-Net [28] proposed to learn to predict the IoU of a detection as the localization confidence and then combine the classification score and localization confidence as the final probability of a detection. Double-Head R-CNN [29] adopts different head architectures for different tasks, i.e., fully connected head for classification and convolution head for regression. In our methods, we aim to improve the classification score by extracting aligned features for each instance. Especially when detecting densely packed objects in aerial images, accurate features are important to robust classification and precise localization. Besides, as discussed in [29], shared features from the backbone are not suitable for both classification and localization. Inspired by [14] and [30], we first adopt ARF to encode the orientation information and then extract orientation-sensitive features and orientation-invariant features for regression and classification, respectively.

III. Proposed Method

In this section, we first enable RetinaNet for oriented object detection and select it as our baseline in Section III-A. Then, we detail the alignment convolution in Section III-B. The architectures of FAM and ODM are presented in Sections III-C and III-D, respectively. Finally, we show details of the proposed S^2A-Net in both training and inference phases. The overall architecture is shown in Fig. 2, and the code is available at https://github.com/csuhan/s2anet.

A. RetinaNet as Baseline

We choose a representative single-shot detector, RetinaNet [12] as our baseline. It consists of a backbone network and two task-specific sub-networks. Feature pyramid network (FPN) [31] is adopted as the backbone network to extract multiscale features. Classification and regression sub-networks are fully convolutional networks with several (i.e., 4) stacked convolution layers. Moreover, Focal loss is proposed to address the extreme foreground–background class imbalance during training.

It is noted that RetinaNet is designed for generic object detection, outputting horizontal bounding box [Fig. 3(a)] represented as

$$\{(\mathbf{x}, w, h)\}$$

with $\mathbf{x} = (x_1, x_2)$ as the center of the bounding box. To be compatible with oriented object detection, we replace the

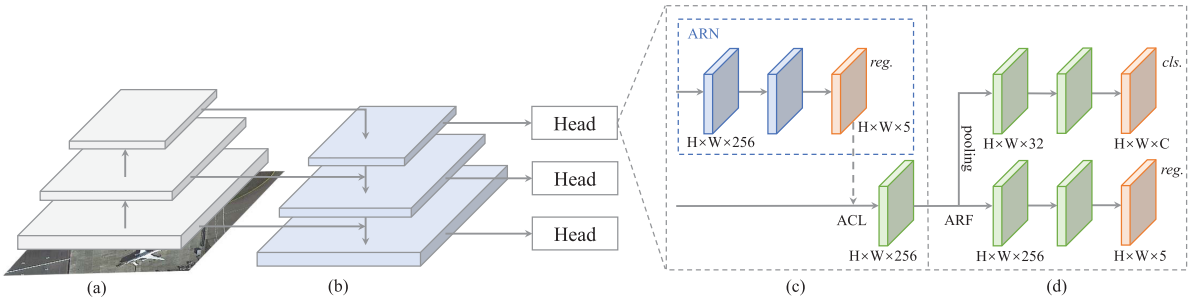


Fig. 2. Architecture of the proposed S^2A-Net. S^2A-Net consists of a backbone network, a feature pyramid network [12], a FAM and an ODM. The FAM and ODM make up the detection head which is applied to each scale of the feature pyramid. In FAM, the ARN is proposed to generate high-quality rotated anchors. Then we feed the anchors and input features into the ACL to extract aligned features. It is noted that we only visualize the regression (*reg.*) branch of ARN and ignore the classification (*cls.*) branch for simplification. In ODM, we first adopt ARF [14] to generate orientation-sensitive features, and pool the features to extract orientation-invariant features. Then the *cls.* branch and *reg.* branch are applied to produce the final detections. (a) Backbone. (b) Feature pyramid network. (c) FAM. (d) ODM.

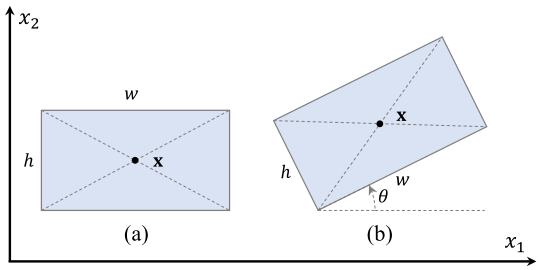


Fig. 3. Two types of bounding box. (a) Horizontal bounding box $\{(\mathbf{x}, w, h)\}$ with center point $\mathbf{x} = (x_1, x_2)$, width w and height h. (b) Oriented bounding box $\{(\mathbf{x}, w, h, \theta)\}$. $\mathbf{x}$ denotes the center point. w and h represent the long side and short side of a bounding box, respectively. θ means the angle from the position direction of x_1 to the direction of w where $\theta \in [-(\pi/4), (3\pi/4)]$. And an oriented bounding box turns to a horizontal one when $\theta = 0$, e.g., $(\mathbf{x}, w, h, 0)$.

regression output of the RetinaNet with oriented bounding box [Fig. 3(b)] as

$$\{(\mathbf{x}, w, h, \theta)\}$$

where $\theta \in [-(\pi/4), (3\pi/4)]$ denotes the angle from the position direction of x_1 to the direction of the width w [4]. All other settings keep unchanged with original RetinaNet.

B. Alignment Convolution

In a standard 2-D convolution, we first sample over the input feature map $\mathbf{X}$ defined on $\Omega = \{0, 1, \ldots, H-1\} \times \{0, 1, \ldots, W-1\}$ by a regular grid $\mathcal{R} = \{(r_x, r_y)\}$, and then sum up the sampled values weighted by $\mathbf{W}$. For example, the grid $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}$ represents a kernel size 3×3 and dilation 1. For each location $\mathbf{p} \in \Omega$ on the output feature map $\mathbf{Y}$, we have

$$\mathbf{Y}(\mathbf{p}) = \sum_{\mathbf{r} \in \mathcal{R}} \mathbf{W}(\mathbf{r}) \cdot \mathbf{X}(\mathbf{p} + \mathbf{r}). \tag{1}$$

Compared with standard convolution, alignment convolution (AlignConv) adds an additional offset field $\mathcal{O}$ for each location $\mathbf{p}$, that is

$$\mathbf{Y}(\mathbf{p}) = \sum_{\mathbf{r} \in \mathcal{R};\, \mathbf{o} \in \mathcal{O}} \mathbf{W}(\mathbf{r}) \cdot \mathbf{X}(\mathbf{p} + \mathbf{r} + \mathbf{o}). \tag{2}$$

As shown in Fig. 4(c) and (d), for location $\mathbf{p}$, the offset field $\mathcal{O}$ is calculated as the difference between anchor-based sampling locations and regular sampling locations (i.e., $\mathbf{p} + \mathbf{r}$). Let $(\mathbf{x}, w, h, \theta)$ represents the corresponding AB at location $\mathbf{p}$. For each $\mathbf{r} \in \mathcal{R}$, the anchor-based sampling location $\mathbf{L}_\mathbf{p}^\mathbf{r}$ can be defined as

$$\mathbf{L}_\mathbf{p}^\mathbf{r} = \frac{1}{S}\left(\mathbf{x} + \frac{1}{k}(w, h) \cdot \mathbf{r}\right) R^T(\theta) \tag{3}$$

where k indicates the kernel size, S denotes the stride of the feature map, and $R(\theta) = (\cos\theta, -\sin\theta;\ \sin\theta, \cos\theta)^T$ is the rotation matrix, respectively. The offset field $\mathcal{O}$ at location $\mathbf{p}$ is

$$\mathcal{O} = \left\{\mathbf{L}_\mathbf{p}^\mathbf{r} - \mathbf{p} - \mathbf{r}\right\}_{\mathbf{r} \in \mathbf{R}}. \tag{4}$$

In this way, we can transform the axis-aligned convolutional features $\mathbf{X}(\mathbf{p})$ of a given location $\mathbf{p}$ into arbitrary-oriented ones based on the corresponding AB.

Comparisons With Other Convolutions: As shown in Fig. 4, standard convolution samples over the feature map by a regular grid. DeformConv learns an offset field to augment the spatial sampling locations. However, it may sample from wrong locations with weak supervision, especially for densely packed objects. Our proposed AlignConv extracts grid-distributed features with the guide of ABs by adding an additional offset field. Different from DeformConv, the offset field in AlignConv is inferred from the ABs directly. The examples in Fig. 4(c) and (d) illustrate that our AlignConv can extract accurate features inside the ABs.

C. Feature Alignment Module

This section introduces the FAM that consists of an anchor refinement network and an alignment convolution layer (ACL) as illustrated in Fig. 2(c).

1) Anchor Refinement Network: The anchor refinement network (ARN) is a light network with two parallel branches: an anchor classification branch (not shown in the figure) and an anchor regression branch. The anchor classification branch classifies anchors into different categories and the anchor regression branch refines horizontal anchors into rotated anchors with high quality. By default, since we only need the

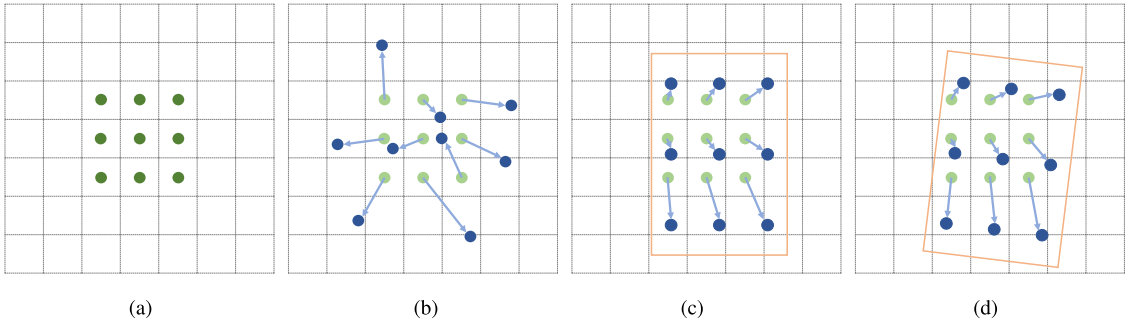


Fig. 4. Illustration of the sampling locations in different methods with 3×3 kernel. (a) the standard 2-D convolution with regular sampling locations (in green dots). (b) Deformable Convolution [25] with deformable sampling locations (in blue dots). (c) and (d) Two examples of our proposed AlignConv with horizontal and rotated AB, respectively (in orange rectangle). The blue arrows mean the offset field.

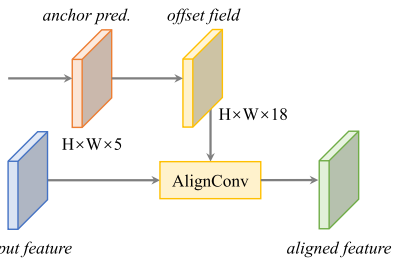


Fig. 5. ACL. It takes the input feature and the anchor prediction (*pred.*) map as inputs and produces aligned features.

regressed ABs to adjust the sampling locations in AlignConv, the classification branch is discarded in the inference phase to speed up the model. But for a fast version of S^{2}A-Net, for which the output of ARN is adopted to produce the final predictions (see Section IV-D), the classification branch is reserved. Following the one-to-one fashion in anchor-free detectors, we preset one squared anchor for each location in the feature map. And we do not filter out the predictions with low confidence because we notice that some negative predictions turn to positive in the final predictions.

2) Alignment Convolution Layer: With AlignConv embedded, we form an ACL which is shown in Fig. 5. Specifically, for each location in the $H \times W \times 5$ anchor prediction map, we first decode it into absolute ABs $(\mathbf{x}, w, h, \theta)$. Then the offset field calculated by (4) along with the input feature is fed into AlignConv to extract aligned features. Note for each AB (5-dimension), we regularly sample 9 (3 rows and 3 columns) points to obtain the 18-dimension offset field [i.e., the x-offset and y-offset of nine points, see the blue arrows in Fig. 4(c) and (d)]. Besides, it should be emphasized that ACL is a light convolution layer with negligible speed latency in offset field calculation.

D. Oriented Detection Module

As shown in Fig. 2(d), the ODM is proposed to alleviate the inconsistency between classification score and localization accuracy and then performs accurate object detection. We first adopt ARF [14] to encode the orientation information. An ARF is a $k \times k \times N$ filter that actively rotates $N-1$ times during convolution to produce a feature map with N orientation channels (N is 8 by default). For a feature map $\mathbf{X}$ and an ARF $\mathbf{F}$, the i-th orientation output of $\mathbf{Y}$ can be denoted as

$$\mathbf{Y}^{(i)} = \sum_{n=0}^{N-1} \mathbf{F}_{\theta_i}^{(n)} \cdot \mathbf{X}^{(n)}, \quad \theta_i = i\frac{2\pi}{N}, \ i = 0, \ldots, N-1 \quad (5)$$

where $\mathbf{F}_{\theta_i}$ is the clockwise θ_i-rotated version of $\mathbf{F}$, and $\mathbf{F}_{\theta_i}^{(n)}$ and $\mathbf{X}^{(n)}$ are the n-th orientation channel of $\mathbf{F}_{\theta_i}$ and $\mathbf{X}$, respectively. Applying ARF to the convolution layer, we can obtain orientation-sensitive features with explicitly encoded orientation informations. The bounding box regression task benefits from the orientation-sensitive features, while the object classification task requires invariant features. Following [14], we aims to extract orientation-invariant features by pooling the orientation-sensitive features. This is simply done by choosing the orientation channel with strongest response as the output feature $\hat{\mathbf{X}}$

$$\hat{\mathbf{X}} = \max \mathbf{X}^{(n)}, \quad 0 < n < N-1. \quad (6)$$

In this way, we can align the feature of objects with different orientations, toward robust object classification. Compared with the orientation-sensitive feature, the orientation-invariant feature is efficient with fewer parameters. For example, an $H \times W \times 256$ feature map with eight orientation channels becomes $H \times W \times 32$ after pooling. Finally, we feed the orientation-sensitive feature and orientation-invariant feature into two sub-networks to regress the bounding boxes and classify the categories, respectively.

E. Single-Shot Alignment Network

We adopt RetinaNet as the baseline, including its network architecture and most parameter settings, and form S^{2}A-Net based on the combination of FAM and ODM. In the following, we detail S^{2}A-Net in both training and inference phases.

1) Regression Targets: Following previous works, we give the parameterized regression targets as

$$\begin{aligned} \Delta\mathbf{x}_g &= (\mathbf{x}_g - \mathbf{x})R(\theta) \cdot \left(\frac{1}{w}, \frac{1}{h}\right) \\ (\Delta w_g, \Delta h_g) &= \log(w_g, h_g) - \log(w, h) \\ \Delta\theta_g &= \frac{1}{\pi}(\theta_g - \theta + k\pi) \end{aligned} \quad (7)$$

where $\mathbf{x}_g$ and $\mathbf{x}$ are for the ground-truth box and the AB, respectively (likewise for w, h, θ). And k is an integer to ensure $(\theta_g - \theta + k\pi) \in [-(\pi/4), (3\pi/4)]$ (see Fig. 3). In FAM, we set $\theta = 0$ to represent a horizontal anchor. Then the regression targets can be expressed by (7). In ODM, we first decode the output of FAM and then re-compute the regression targets by (7).

2) Matching Strategy: We adopt IoU as the metrics, and an AB can be assigned to positive (or negative) if its IoU is greater than a foreground threshold (or less than a background threshold, respectively). Different from the IoU between horizontal bounding boxes, we calculate the IoU between two oriented bounding boxes. By default, we set the foreground threshold as 0.5 and the background threshold as 0.4 in both FAM and ODM.

3) Loss Function: The loss of S^{2}A-Net is a multitask one which consists of two parts, i.e., the loss of FAM and the loss of ODM. For each part, we assign a class label to each anchor/refined anchor and regress its location. The loss function can be defined as

$$\mathcal{L} = \frac{1}{N_F}\left(\sum_i \mathcal{L}_c\left(c_i^F, l_i^*\right) + \sum_i \mathbf{1}_{\left[l_i^* \geq 1\right]} \mathcal{L}_r\left(\mathbf{x}_i^F, \mathbf{g}_i^*\right)\right) + \frac{\lambda}{N_O}\left(\sum_i \mathcal{L}_c\left(c_i^O, l_i^*\right) + \sum_i \mathbf{1}_{\left[l_i^* \geq 1\right]} \mathcal{L}_r\left(\mathbf{x}_i^O, \mathbf{g}_i^*\right)\right) \quad (8)$$

where λ is a loss balance parameter, $\mathbf{1}_{[\cdot]}$ is an indicator function, N_F and N_O are the numbers of positive samples in the FAM and ODM, respectively, i is the index of a sample in a mini-batch. c_i^F and $\mathbf{x}_i^F$ are the predicted category and refined locations of the anchor i in FAM. c_i^O and $\mathbf{x}_i^O$ are the predicted object category and locations of the bounding box in ODM. l_i^* and $\mathbf{g}_i^*$ are the groundtruth category and locations of the anchor i. The focal loss [12] and smooth $L1$ loss are adopted as the classification loss $\mathcal{L}_c$ and the regression loss $\mathcal{L}_r$, respectively.

4) Inference: S^{2}A-Net is a fully convolutional network and we can simply forward an image through the network without complex RoI operation. Specifically, we pass the input image to the backbone network to extract pyramid features. Then the pyramid features are fed into FAM to produce refined anchors and aligned features. After that, ODM encodes the orientation information to produce the predictions with high confidence. Finally, we choose top-k (i.e., 2000) predictions and adopt NMS to yield the final detections.

IV. Experiments and Analysis

A. Data Sets

1) DOTA [3]: It is a large aerial image data set for oriented objects detection which contains 2806 images with the size ranges from 800×800 to 4000×4000 and 188 282 instances of 15 common object categories includes: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC).

TABLE I
Results of Different RetinaNet on DOTA. Depth Indicates the Number of Convolution Layer in Two Subnetworks of RetinaNet

	Model	#Anchor	Depth	mAP	GFLOPs	Param
(a)	RetinaNet	9	4	68.05	215.92	36.42 M
(b)	RetinaNet	9	2	67.64	164.38	34.06 M
(c)	RetinaNet	1	2	67.00	156.33	33.69 M

Both training and validation sets are used for training, and the testing set is used for testing. Following [3], we crop a series of 1024×1024 patches from original images with a stride of 824. We only adopt random horizontal flipping during training to avoid over-fitting and no other tricks are utilized if not specified. For fair comparison with other methods, we adopt data augmentation (i.e., random rotation) in the training phase. For multiscale experiments, we firstly resize original images at three scales (0.5, 1.0 and 1.5) and then crop them into 1024×1024 patches with a stride of 512.

2) HRSC2016 [15]: It is a high-resolution ship recognition data set annotated with oriented bounding boxes which contains 1061 images, and the image size ranges from 300×300 to 1500×900. We use the training (436 images) and validation (181 images) sets for training and the testing set (444 images) for testing. All images are resized to $(800, 512)$ without changing the aspect ratio. Horizontal flipping is applied during training.

B. Implementation Details

We adopt ResNet101 FPN as the backbone network for fair comparison with other methods, and ResNet50 FPN is adopted for other experiments if not specified. For each level of pyramid features (i.e., P_3 to P_7), we preset one squared anchor per location with a scale of four times the total stride size (i.e., 32, 64, 128, 256, 512). The loss balance parameter λ is set to 1. The hyperparameters of Focal loss $\mathcal{L}_c$ are set to $\alpha = 0.25$ and $\gamma = 2.0$. We adopt the same training schedules as mmdetection [32]. We train all models in 12 epochs for DOTA and 36 epochs for HRSC2016. SGD optimizer is adopted with an initial learning rate of 0.01 and the learning rate is divided by 10 at each decay step. The momentum and weight decay are 0.9 and 0.0001, respectively. We adopt learning rate warm-up for 500 iterations. We use four V100 GPUs with a total batch size of eight for training and a single V100 GPU for inference by default. The time of post-processing (e.g., NMS) is included in all experiments.

C. Ablation Studies

In this section, we conduct a series of experiments on the testing set of DOTA to validate the effectiveness of our method. ResNet50 FPN is adopted as the backbone in all experiments. It isi noted that we extend the `flops_counter` tool in mmdetection [32] to calculate the FLOPs of our method.

1) RetinaNet as Baseline: As a single-shot detector, RetinaNet is fast enough. However, any module added to it will introduce more computations. We experiment different architectures and settings on RetinaNet. As shown in Table I(a),

TABLE II

COMPARING ALIGNMENT CONVOLUTION (ALIGNCONV) WITH OTHER CONVOLUTION METHODS. WE COMPARE OUR ALIGNCONV WITH THE STANDARD CONVOLUTION (CONV), DEFORMABLE CONVOLUTION (DEFORMCONV), AND GUIDED ANCHORING DEFORMABLE CONVOLUTION (GA-DEFORMCONV)

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	GFLOPs
Conv	88.87	76.34	46.42	67.53	77.21	74.80	82.27	90.79	81.22	85.02	50.99	61.10	63.54	67.24	53.25	71.11	**196.62**
DeformConv	88.96	80.23	45.92	67.51	77.10	74.23	84.28	90.81	81.47	85.56	54.19	**64.11**	64.85	68.13	48.34	71.71	198.02
GA-DeformConv	88.72	79.56	46.19	65.41	76.86	74.96	79.44	90.78	80.99	84.73	55.31	63.17	62.07	67.69	54.12	71.33	197.92
AlignConv	**89.11**	**82.84**	**48.37**	**71.11**	**78.11**	**78.39**	**87.25**	**90.83**	**84.90**	**85.64**	**60.36**	62.60	**65.26**	**69.13**	**57.94**	**74.12**	198.03

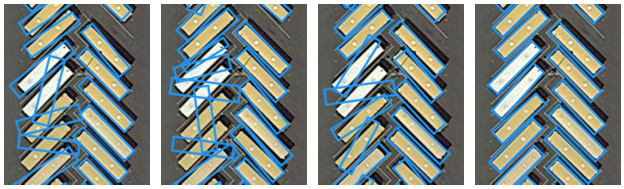


Fig. 6. Qualitative comparison of different convolution methods. The blue bounding box indicates the prediction of large vehicle.

TABLE III

ABLATION STUDIES. WE CHOOSE A LIGHT RETINANET [SHOWN IN TABLE I(c)] AS THE BASELINE, AND EXPERIMENT DIFFERENT SETTINGS OF S^{2}A-NET, I.E., ARN, ACL, AND ARF

	Baseline	Different Settings of S^{2}A-Net					
ARN			✓		✓	✓	✓
ACL						✓	✓
ARF				✓	✓		✓
mAP	67.00	68.26	71.17	68.35	71.11	73.24	74.12

TABLE IV

EXPERIMENTS OF DIFFERENT NETWORK DESIGNS. WE EXPLORE THE NETWORK DESIGN IN FAM AND ODM WITH DIFFERENT NUMBER OF LAYERS. SETTING (d) IS THE DEFAULT SETTING OF OUR PROPOSED METHOD SHOWN IN FIG. 2

	Model	FAM	ODM	mAP	GFLOPs	Param
(a)	RetinaNet	-	-	68.05	215.92	36.42 M
(b)	S^{2}A-Net	1	1	73.04	**159.27**	**33.25 M**
(c)	S^{2}A-Net	1	3	72.89	210.81	35.61 M
(d)	S^{2}A-Net	2	2	**74.12**	198.03	35.02 M
(e)	S^{2}A-Net	1	3	72.86	185.04	34.43 M
(f)	S^{2}A-Net	4	4	73.30	275.22	38.57 M

RetinaNet achieves a mAP of 68.05% with 215.92 GFLOPs and 36.42 M parameters, indicating that our baseline is solid. If the depth of RetinaNet head changes from 4 to 2, the mAP drops 0.41% and the FLOPs (*resp.* parameters) reduce 51.54 G (*resp.* 2.36 M). Furthermore, if we set one anchor per location [Table I(c)], the FLOPs reduces 28% with a accuracy drop of 1.5% compared with Table I(a). The results show that a light detection head and few anchors can also achieve competitive performance and better speed–accuracy tradeoff.

2) Effectiveness of AlignConv: As discussed in Section III-B, we compare AlignConv with other methods to validate its effectiveness. We only replace AlignConv with other convolution methods and keep other settings unchanged. Besides, we also add comparison with Guided Anchoring DeformConv (GA-DeformConv) [26]. It is noted that the offset field of GA-DeformConv is learned from the anchor prediction map in ARN by a 1×1 convolution.

(a) (b)

Fig. 7. Qualitative comparison of detection results. We crop a large-size image into 1024×1024 chip images with a stride of 824. The large-size image and chip images are fed into the same network to produce detection results (e.g., planes in red boxes) without resizing. Instances with the same number are corresponding. (a) Detection on chip images. (b) Detection on large-size image.

As shown in Table II, AlignConv surpasses other methods by a big margin. Compared with the standard convolution, AlignConv improves about 3% mAP while only introduces 1.41 GFLOPs computation. Besides, AlignConv improves the performance for almost all categories, especially for those categories with large aspect ratios (e.g., bridge), densely distribution (e.g., SVs and large vehicles), and fewer instances (e.g., helicopters). On the contrary, DeformConv and GA-DeformConv only achieve 71.71% and 71.33% mAP, respectively. The qualitative comparison in Fig. 6 shows that AlignConv achieves accurate bounding box regression in detecting densely packed and arbitrary-oriented objects, while other methods with implicit learning get poor performance.

3) Effectiveness of ARN and ARF: To evaluate the effectiveness of ARN and ARF, we experiment different settings of S^{2}A-Net. If ARN is discarded, then FAM and ODM share the same initial anchors without refinement. If ARF is discarded, we replace the ARF layer with the standard convolution layer. As shown in Table III, without ARN, ACL, and ARF, our method achieves 68.26% mAP, about 1.26% mAP higher than the baseline method. This is mainly because we add supervisions in both FAM and ODM. With the participation of ARN, we obtain 71.17% mAP, showing that anchor refinement is important to the final predictions in ODM.

Besides, we find ARF dose nothing for performance improvement without the participation of ACL, i.e., applying ARF or the combination of ARN and ARF to our method only achieve 68.35% and 71.11% mAP, respectively. However,

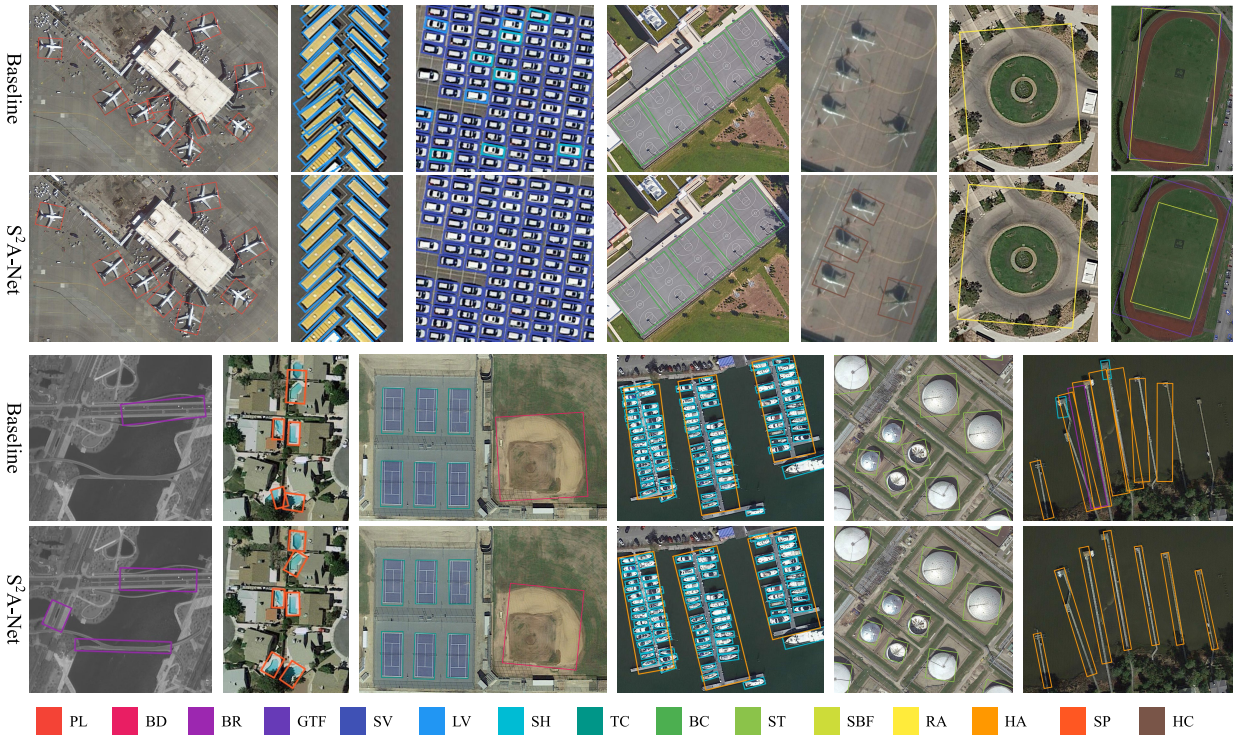


Fig. 8. Some detection results on DOTA by different methods. For each image pair, the upper is the baseline method while the bottom is by S^2A-Net.

if we put ACL and ARF together, there is an obvious improvement, from 73.24% to 74.12%. We argue that CNNs are not rotation-invariant, and even we can extract accurate features to represent the object, the corresponding features are still rotation-sensitive. So the participation of ARF augments the orientation information explicitly, leading to better regression and classification.

4) Network Design: As shown in Table IV, we explore different network designs in FAM and ODM. Compared with the baseline method in Table IV(a), we can conclude that S^2A-Net is not only an effective detector with high detection accuracy, but also an efficient detector in both speed and parameters. The results in Table IV(b)–(f) show that our proposed method is insensitive to the depth of the network and the performance improvements mainly come from our novel alignment mechanism. Besides, as the number of layers increases, there is a performance drop from Table IV(d)–(f). We hypothesize that deeper networks with a larger receptive field may hinder the detection performance of small size objects. Moreover, the setting (d), for which the number of layers in FAM and ODM is the same, obtains the highest mAP among (c)–(e), while (c) and (e) have a significant drop in mAP, showing that similar receptive field in FAM and ODM is more balancing for high-quality object detection.

TABLE V

COMPARISON OF DIFFERENT SETTINGS DETECTING ON LARGE IMAGES IN DOTA. STRIDE IS THE CROPPING STRIDE REFERRED IN SECTION IV-A. **#Image** MEANS THE NUMBER OF IMAGES OR CHIPS. OUTPUT INDICATES THE MODULE (I.E., FAM OR ODM) USED FOR TESTING. WE SHOW THE INFERENCE TIME REQUIRED FOR ENTIRE DATA SET USING *FP32*/*FP16* WITH FOUR V100 GPUS

Input Size	Stride	#Image	Output	mAP	Time (s)
1024 × 1024	1024	8143	ODM	71.20	150 / 126
1024 × 1024	824	10833	ODM	74.12	246 / 160
1024 × 1024	512	20012	ODM	74.62	352 / 308
Original	-	937	ODM	74.01	120 / 103
Original	-	937	FAM	70.85	104 / 97

TABLE VI

COMPARING S^2A-NET WITH CLUSDET [33] ON DOTA VALIDATION SET. FOLLOWING [33], WE REPORT THE ACCURACY OF FIVE CATEGORIES (I.E., PL, SV, LV, SH AND HC) WITH DIFFERENT IOU THRESHOLDS (I.E., $mAP_{.5}$, $mAP_{.75}$ AND $mAP_{.5:.95}$). THE RESULTS OF RETINANET AND S^2A-NET ARE CALCULATED FROM THE AXIS-ALIGNED BOUNDING BOXES OF THE OUTPUT. #IMAGE MEANS THE NUMBER OF IMAGES OR CHIPS. † INDICATES THAT THE OUTPUT OF FAM IS ADOPTED FOR THE FINAL RESULTS

Methods	#Image	$mAP_{.5:.95}$	$mAP_{.5}$	$mAP_{.75}$
ClusDet [33]	1055	32.2	47.6	39.2
RetinaNet	**458**	41.6	70.5	44.2
S^2A-Net† (Ours)	**458**	42.7	72.7	45.3
S^2A-Net (Ours)	**458**	**43.9**	**75.8**	**46.3**

D. Detecting on Large-Size Images

The size of aerial image often ranges from thousands to tens of thousands, which means more computations and memory footprint. Many previous works [3], [4] adopt a detection on chips strategy to alleviate this challenge, even if a chip does not contain any object. ClusDet [33] tries to address this issue by generating clustered chips, while introducing more complex operations (e.g., chip generation and results merge) and significant performance drop. As our proposed S^2A-Net is efficient and the architecture is flexible, we aims to detect objects on large-size images directly.

TABLE VII

COMPARISONS WITH STATE-OF-THE-ART METHODS ON DOTA. R-101-FPN STANDS FOR RESNET 101 WITH FPN (LIKEWISE R-50-FPN), AND H-104 STANDS FOR HOURGLASS 104. † INDICATES TRAINING AND TESTING WITHOUT DATA AUGMENTATION. ‡ DENOTES THE INPUT IS THE ORIGINAL IMAGES OTHER THAN CHIP IMAGES. * MEANS MULTISCALE TRAINING AND TESTING

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	FPS
two-stage:																		
FR-O [3]	R-101	79.42	77.13	17.70	64.05	35.30	38.02	37.16	89.41	69.64	59.28	50.30	52.91	47.89	47.40	46.30	54.13	-
Azimi *et al.* [34]	R-101-FPN	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16	-
RoI Trans.* [4]	R-101-FPN	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56	5.9
CADNet [5]	R-101-FPN	87.80	82.40	49.40	73.50	71.10	63.50	76.60	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90	-
SCRDet [8]	R-101-FPN	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21	72.61	-
Xu *et al.* [7]	R-101-FPN	89.64	**85.00**	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	**70.91**	72.94	70.86	57.32	75.02	10.0
CenterMap-Net [6]	R-50-FPN	88.88	81.24	53.15	60.65	78.62	66.55	78.10	88.83	77.80	83.61	49.36	66.19	72.10	72.36	58.70	71.74	-
CenterMap-Net* [6]	R-101-FPN	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03	-
one-stage:																		
RetinaNet [12]	R-101-FPN	88.82	81.74	44.44	65.72	67.11	55.82	72.77	90.55	82.83	76.30	54.19	63.64	63.71	69.73	53.37	68.72	12.7
DRN [35]	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70	-
DRN* [35]	H-104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23	-
R^3Det [24]	R-101-FPN	89.54	81.99	48.46	62.52	70.48	74.29	77.54	90.80	81.39	83.54	61.97	59.82	65.44	67.46	60.05	71.69	-
R^3Det [24]	R-152-FPN	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74	-
S^2A-Net† (Ours)	R-50-FPN	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12	16.0
S^2A-Net†‡ (Ours)	R-50-FPN	89.11	81.51	48.75	72.85	78.23	76.77	86.95	90.84	83.59	85.52	62.70	61.63	66.55	68.94	56.24	74.01	**22.6**
S^2A-Net (Ours)	R-101-FPN	88.70	81.41	54.28	69.75	78.04	80.54	88.04	90.69	84.75	86.22	65.03	65.81	76.16	73.37	58.86	76.11	12.7
S^2A-Net* (Ours)	R-50-FPN	88.89	83.60	**57.74**	**81.95**	79.94	**83.19**	89.11	90.78	84.87	**87.81**	70.30	68.25	78.30	77.01	**69.58**	**79.42**	16.0
S^2A-Net* (Ours)	R-101-FPN	89.28	84.11	56.95	79.21	**80.18**	82.93	**89.21**	90.86	84.66	87.61	**71.66**	68.23	**78.58**	**78.20**	65.55	79.15	12.7

TABLE VIII

COMPARISONS OF STATE-OF-THE-ART METHODS ON HRSC2016. #ANCHOR MEANS THE NUMBER OF ANCHORS AT EACH LOCATION OF THE FEATURE MAP. * INDICATES THAT THE RESULT IS EVALUATED UNDER PASCAL VOC2012 METRICS

Methods	RC2 [36]	R^2PN* [37]	RRD [30]	RoI Trans. [4]	Xu *et al.* [7]	R^3Det [24]	DRN [35]	CenterMap-Net [6]	S^2A-Net (Ours)
#Anchor	-	24	13	20	20	21	-	15	**1**
mAP	75.7	79.6	84.3	86.2	88.2	89.26	92.7*	92.8*	**90.17 / 95.01***

We first explore different settings of the input size and cropping stride, and report the mAP and overall time during inference (Table V). We first crop the images into 1024 × 1024 chips, and the mAP improves from 71.20% to 74.62% when the stride decreases from 1024 to 512. However, the number of chip images increases from 8143 to 20012, and the overall inference time increases about 135%. If we detect on the original large-size images without cropping, the inference time has reduced by 50% with negligible loss of accuracy. We argue that the cropping strategy makes it hard to detect objects around the boundary (Fig. 7). Besides, if we adopt the output of FAM for detection and Floating-Point 16 (FP16) to speed up the inference, we can reduce the inference time to 97 seconds with a mAP of 70.85%. Compared our S^2A-Net with ClusDet [33] (Table VI), our method only process 458 images and outperforms ClusDet by a large margin. If we adopt the output of FAM for evaluation, we still achieve 42.7% $mAP_{.5-.95}$ and 72.7% $mAP_{.5}$. The result demonstrates that out method is efficient and effective, and our detection strategy can achieve better speed–accuracy tradeoff.

E. Comparisons With the State-of-the-Art

In this section, we compare our proposed S^2A-Net with other state-of-the-art methods on two aerial detection data sets,:DOTA and HRSC2016. The settings have been introduced in Sections IV-A and IV-B.

1) Results on DOTA[2]*:* It is noted that RetinaNet is our re-implemented version referred in Section III-A. As shown in Table VII, we achieve 74.01% mAP in 22.6 frames per second (FPS) with ResNet-50-FPN backbone and without any data augmentation (e.g., random rotation). It is noted that the FPS is an average FPS and we obtain it by calculating the overall inference time and the number of chip images (i.e., 10 833). Besides, we achieve state-of-the-art 76.11% mAP with a ResNet101 FPN backbone, outperforming all two-stage and one-stage methods. In multiscale experiments, our S^2A-Net achieves 79.42% and 79.15% mAP with a ResNet-50-FPN and ResNet-101-FPN backbone, respectively. And we achieve best results in 10/15 categories, especially for some hard categories (e.g., bridge, SBF, SP, helicopter). Qualitative detection results of the baseline method (i.e., RetinaNet) and our S^2A-Net are visualized in Fig. 8. Compared with RetinaNet, our S^2A-Net produces less false predictions when detecting on the object with dense distribution and large-scale variations.

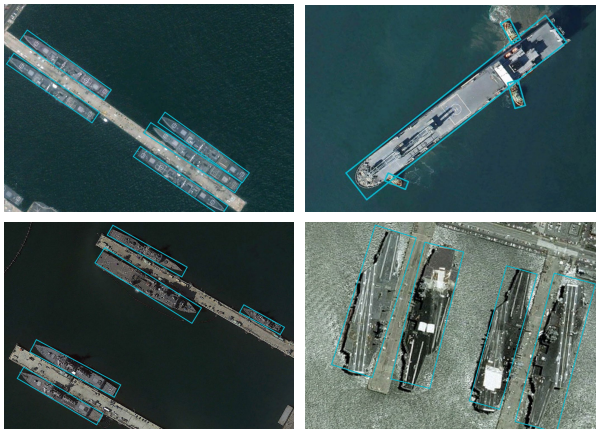


Fig. 9. Some detection results on HRSC2016 with the proposed S^2A-Net.

[2]The result is available at https://captain-whu.github.io/DOTA/results.html with setting name `hanjiaming`. It is noted that to concentrate on studying the algorithmic problem of ODAI, this setting is without using model fusions which can further improve the detection performance.

2) Results on HRSC2016: It is noted that DRN [35] and CenterMap-Net [6] are evaluated under PASCAL VOC2012 metrics while other methods are evaluated under PASCAL VOC2007 metrics, and the performance under VOC2012 metrics is better than that under VOC2007 metrics. As shown in Table VIII, our proposed S^2A-Net achieves 90.17% and 95.01% mAP under VOC2007 and VOC2012 metrics, respectively, outperforming all other methods. The objects in HRSC2016 have large aspect ratios and arbitrary orientations, and previous methods often set more anchors for better performance, e.g., 20 in RoI Trans. and 21 in R^3Det. Compared with the previous best result 89.26% (VOC2007) by R^3Det and 92.8% (VOC2012) by CenterMap-Net, we improve 0.91% and 2.21% mAP respectively with only one anchor, which effectively get rid of heuristically defined anchors. Some qualitative results are shown in Fig. 9.

V. Conclusion

In this article, we propose a simple and effective single-shot alignment network (S^2A-Net) for oriented object detection in aerial images. With the proposed FAM and ODM, our S^2A-Net realizes full feature alignment and alleviates the inconsistency between regression and classification. Besides, we explore the approach to detect on large-size images for better speed-accuracy trade-off. Extensive experiments demonstrate that our S^2A-Net can achieve state-of-the-art performance on both DOTA and HRSC2016.

References

[1] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[2] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[3] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[4] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.

[5] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[6] J. Wang, W. Yang, H.-C. Li, H. Zhang, and G.-S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 28, 2020, doi: 10.1109/TGRS.2020.3010051.

[7] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 18, 2020, doi: 10.1109/TPAMI.2020.2974745.

[8] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[13] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4969–4978.

[14] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4961–4970.

[15] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[18] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[19] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. ECCV*, 2018, pp. 734–750.

[20] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[21] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9656–9665.

[22] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 900–904.

[23] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[24] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*. [Online]. Available: http://arxiv.org/abs/1908.05612

[25] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[26] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2960–2969.

[27] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," 2019, *arXiv:1908.01570*. [Online]. Available: http://arxiv.org/abs/1908.01570

[28] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. ECCV*, 2018, pp. 784–799.

[29] Y. Wu *et al.*, "Rethinking classification and localization for object detection," 2019, *arXiv:1904.06493*. [Online]. Available: http://arxiv.org/abs/1904.06493

[30] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.

[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[32] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*. [Online]. Available: http://arxiv.org/abs/1906.07155

[33] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8310–8319.

[34] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. ACCV*, 2018, pp. 150–165.

[35] X. Pan *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11207–11216.

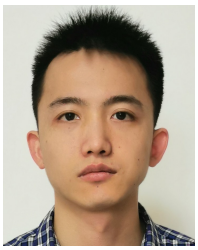
[36] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*. [Online]. Available: http://arxiv.org/abs/1711.09405

[37] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[39] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[40] J. Ding *et al.*, "ICPR2018 contest on object detection in aerial images (ODAI-18)," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1–6.
[41] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 30, 2020, doi: 10.1109/TGRS.2020.3010106.
[42] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[43] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. NIPS*, 2016, pp. 4905–4913.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. ICCV*, 2017, pp. 5562–5570.
[46] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[47] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[48] T. X. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," in *Proc. NIPS*, 2019, pp. 1–11.
[49] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

Jiaming Han received the B.S. degree in remote sensing science and technology from Central South University, Changsha, China, in 2019. He is pursuing the master's degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include object detection and remote sensing.



Jian Ding received the B.S. degree in aircraft design and engineering from Northwestern Polytechnical University, Xian, China, in 2017. He is pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include object detection, instance segmentation and remote sensing.

Jie Li received the Ph.D. degree in image processing and computer vision from Wuhan University, Wuhan, China, in 2017.

From 2016 to 2017, he has been a Visiting Scholar with University of Wisconsin-Milwaukee, Milwaukee, USA, for six months. He is a Researcher with Shanghai Aerospace Electronic Technology Institute, Shanghai, China. His research interests include object detection and remote sensing.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he has been a Post-Doctoral Researcher with the Center de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris. He is working as a Full Professor in computer vision and photogrammetry with Wuhan University. He has also been working as a Visiting Scholar with DMA, École Normale Supérieure (ENS-Paris) in 2018. He is also a Guest Professor with the Future Lab AI4EO in Technical University of Munich (TUM). His research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote-sensing image understanding.

Dr. Xia serves on the Editorial Boards of the Journals *Pattern Recognition*, *Signal Processing: Image Communications*, *EURASIP Journal on Image & Video Processing*, *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.